Диброва К.С., студент
Гольдштейн С.Л., проф., д-р техн. наук

МОДЕРНИЗАЦИЯ СИСТЕМЫ АВТОМАТИЧЕСКОГО ПОСТРОЕНИЯ СПЕЦИАЛИЗИРОВАННОГО ТЕЗАУРУСА

Разработаны сотни тезаурусов, описывающих понятийные и терминологические системы многих предметных областей. Однако разработка тезауруса для новой предметной области, равно как и его пополнение, все еще остается большой проблемой в силу трудоемкости ручной работы.

При автоматизации сложных интеллектуальных процессов важным является не только разработка автоматических процедур, но и:

- обоснованность применения каждой из этих процедур в конкретном случае;
- последовательность их применения;
- относительный «вес» результатов применения каждой процедуры, позволяющий выбрать правильный (из ряда неодинаковых) в случае использования нескольких алгоритмов одновременно, а также
- автоматически контролируемые действия эксперта.

Можно составить кортежную модель качества создаваемого тезауруса при применении автоматических процедур обработки текста:

АСТ = <ОПП, ППП, ВРП, КДЭ>,

где КСТ – адекватность составленного тезауруса; ОПП – обоснованность применения процедуры; ППП – последовательность применения процедур; ВРП – «вес» результатов применения процедур; КДЭ – контроль действий эксперта (при интерпретации полученных данных).

Пути разрешения проблемной ситуации: аналогами этой системы являются программы анализа и лингвистической обработки текстов: TextAnalyst, Mystem, LingSoft. В этих продуктах не реализована поддержка некоторых популярных текстовых форматов, таких как doc (MS Word), PDF (Adobe Reader). Также морфологические анализаторы, например mystem, имеют высокую степень неоднозначности разбора. В связи с этим необходимо ее уменьшить.

Направленность модернизации: систему анализа текстов можно улучшить путем поддержки дополнительных форматов текстовой информации и понизить коэффициент неоднозначности разбора. Для составления тезауруса необходимо использовать специфические словари, подходящие для данной предметной области, а также необходимо учитывать специфику лингвистических закономерностей построения текста по этой предметной области.

Направленность функционирования системы: автоматическое создание тезаурусов, наиболее адекватных и наиболее удовлетворяющих предъявляемым требованиям.

Цель создания системы: значительная экономия человеко-часов на ручное составление тезаурусов понятий заданной предметной области, ускорение обработки текстовой информации для СОЗ, развитие программ анализа и

лингвистической обработки текстов и систем обработки естественного языка, а также наработка тезаурусов для различных предметных областей.